Sentiment Analysis on Amazon Electronics Products Review Dataset

Binay Oli, Hao Zhang, Saheb Singh Johar1

1 Harrisburg University

Author note

The authors made the following contributions. Binay Oli, Hao Zhang, Saheb Singh Johar: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing.

Special Thanks to Dr. Ying Lin for his support throughout the project.

Presentation Video Link - [CISC 520 Presentation Final +recordings V2](https://www.youtube.com/watch?v=TMhs26prb38)

Correspondence concerning this article should be addressed to Binay Oli, Hao Zhang, Saheb Singh Johar, 326 Market St. Harrisburg, PA 17101. E-mail: [boli1@my.harrisburgu.edu](mailto:boli1@my.harrisburgu.edu), h[zhang29@my.harrisburgu.edu](mailto:zhang@my.harrisburgu.edu), sjohar@my.harrisburgu.edu

Abstract

The project “Sentiment Analysis of Amazon Reviews” was aimed at analyzing customer sentiment expressed in Amazon reviews using machine learning techniques. A diverse dataset of Amazon reviews across various product categories was collected, and thorough preprocessing and cleaning were performed to prepare the data for analysis. By leveraging machine learning algorithms, a sentiment analysis model was developed to automatically classify the sentiment of the reviews as positive, negative, or neutral. The sentiment of the reviews was accurately predicted for approximately 62.5% and 87% of the reviews by the logistic regression model and DistillBert model, respectively, achieving accuracy scores of 0.625 and 0.87. Valuable insights into customer sentiment were provided by the project, with positive and negative sentiment trends being identified, and correlations between sentiment and various review attributes being uncovered. Future work includes the improvement of the model’s accuracy, the expansion of analysis to other e-commerce platforms, the incorporation of aspect-based sentiment analysis, the implementation of real-time sentiment monitoring, and the integration of sentiment insights into business decision-making processes.

*Keywords:* Sentiment Analysis, Amazon Reviews, Customer Sentiment, Machine Learning Techniques, Diverse Dataset, Preprocessing, Cleaning, Sentiment Classification, Logistic Regression Model, DistillBert Model, Accuracy Score

*Word count:* 176

Sentiment Analysis on Amazon Electronics Products Review Dataset

# Introduction

In today’s digital age, especially in the post-COVID era, online shopping has become the norm for many people. As the COVID-19 pandemic may be largely behind us, people have transitioned back to their normal daily life. One thing that didn’t change is the way we shopped - more and more relied on online. More and more people relied on online reviews to make purchasing decisions. There have been vast amounts of new products introduced daily on the Amazon platform. It was not easy to go through either the products or the reviews. We looked through different reviews to determine the best product that we wanted to purchase. It was essential to understand the sentiment of the vast reviews on a given product. The goal of this project was to leverage different machine learning algorithms to classify the sentiment of reviews on products so it would be efficient for customers to make purchase decisions.

# Benefits of Sentiment Analysis

Customer feedback: Sentiment analysis allows to gauge the overall sentiment expressed by customers in their reviews and ratings. By analyzing the sentiment, sellers can identify whether customers generally have positive or negative experiences with the products.

Product improvement: Understanding the sentiment associated with specific products helps in identifying areas for improvement. If sentiment analysis reveals recurring negative sentiments, it indicates potential issues or shortcomings that need to be addressed.

Competitor analysis: Analyzing sentiment for Amazon products allows to compare the sentiment of different products within the same category or market segment. This helps in benchmarking against competitors and identifying areas where products excel or lag behind. By understanding customer sentiment towards competing products, sellers can gain insights into their strengths and weaknesses, informing their product development and marketing strategies.

Reputation management: Sentiment analysis provides insights into the overall reputation of products and brands. By tracking sentiment over time, sellers can identify trends, spot potential reputation crises, and take proactive measures to manage and improve brand perception. Positive sentiment can be leveraged for marketing purposes, while negative sentiment can be addressed to protect the brand’s image and reputation.

Decision-making: Sentiment analysis acts as a valuable decision-making tool. It allows sellers to make data-driven decisions based on customer sentiment and feedback. For example, if sentiment analysis reveals consistently positive sentiment for a particular product, it can inform decisions to increase inventory, invest in marketing campaigns, or expand the product line. On the other hand, if sentiment analysis shows negative sentiment, it may trigger actions such as product recalls, quality control measures, or customer service improvements.

# Methods

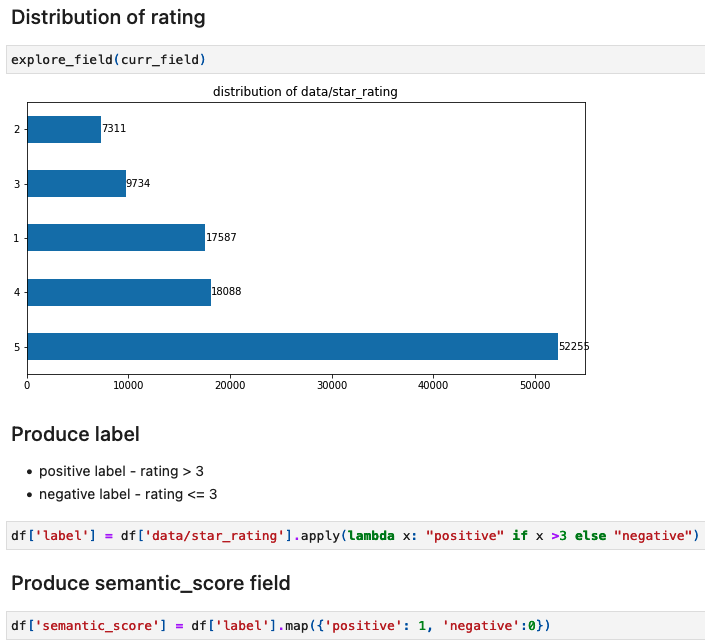
## Methodology For Machine Learning Algorithm

### Data Collection.

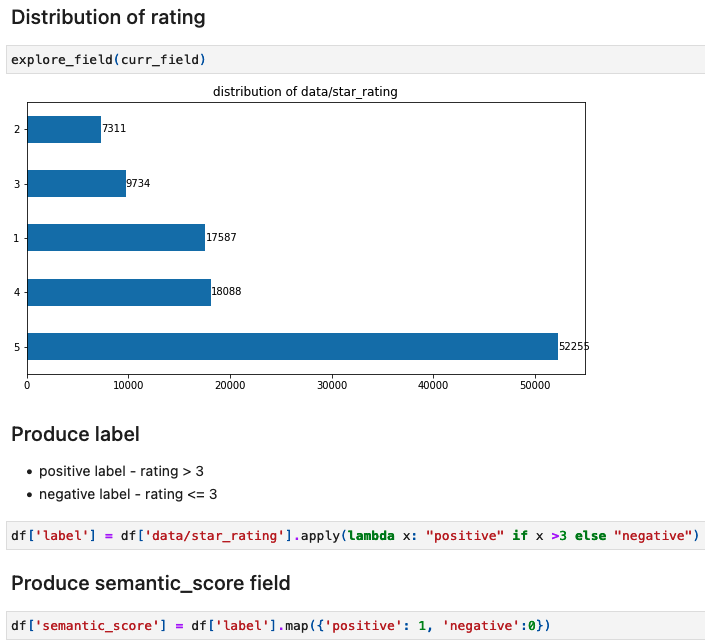
Electronics are very popular on all e-commerce platforms. This is the area of our current focus. The Amazon Electronics Review Dataset from TensorFlow dataset was used for this project. The overall project plan was to clean up the review dataset, apply machine learning and deep learning algorithms to perform sentiment analysis and classify reviews for electronic products.

### EDA on Target variable.

First, an extensive EDA (Exploratory Data Analysis) was employed to discover the features within the dataset. Rating field is the target variable in this project. A visualization of the entire dataset is presented below.



Researching the ratings online, it made sense to classify any rating that was greater or equal to 3 as positive and any rating less than 3 as negative.



Then, we mapped the positive label as 1 and the negative label as 0. The dataset was imbalanced and proved to be very computationally expensive for fine-tuning deep learning algorithms.

To alleviate the computational constraint, the data was cleaned up to create a well-balanced dataset while preserving all the features. During the data exploration analysis, 2000 records of positive reviews and 2000 records of negative reviews were generated as input.

### EDA on Features.

Prior to conducting the sentiment analysis, the dataset underwent a series of preprocessing steps to ensure the cleanliness and consistency of the data. The NLP (Natural Language Processing) pipeline included removing duplicates, handling missing values, and applying text normalization techniques such as lowercasing, punctuation removal, stopword removal, and tokenization.

### Encode Features to Numerical Presentations.

To represent the textual data in a format suitable for the models, feature extraction techniques were employed. For the Logistic Regression model, the feature extraction was performed using the TF-IDF (Term Frequency-Inverse Document Frequency) approach. This method considered the term frequency (TF) of a term in a document and the inverse document frequency (IDF), which penalized common terms across the entire corpus. Categorical data were converted to a numerical representation for the logistic regression to be trained on.

## Methodology For Deep Learning Algorithm

The initial approach employed the widely recognized machine learning library, scikit-learn. For comparison, the second methodology made substantial use of the TensorFlow v2 library. Several crucial steps were involved in fine-tuning the Pretrained Transformer family model. Selecting an appropriate learning rate was one of the key metrics to consider. Generally, for transformer models, a learning rate on the order of 1E-05 was advised. In our project, we found a learning rate of 5E-05 to be the most suitable. We tested both the Adam Optimizer and the Adamax Optimizer, with each exhibiting commendable performance. A batch size of 32 was utilized, and our model was fine-tuned over two epochs on our review dataset.

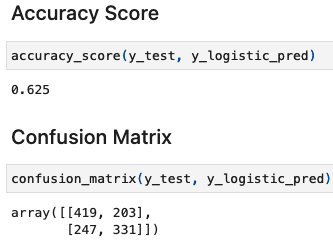
## Data analysis

We used R (Version 4.1.2; R Core Team, 2021) and the R-packages *papaja* (Version 0.1.0.9999; Aust & Barth, 2022), and *tinylabels* (Version 0.2.3; Barth, 2022) for all our analyses.

### Model 1: Logistic Regression Model.

For this model, the best practices of Natural Language Processing are followed.

● Lowercasing - all the text was converted to lowercase to ensure that the model treats words with different capitalization the same. This step prevents the model from considering “Good” and “good” as different words. ● Tokenization - This step helps in breaking down the text into smaller units, enabling further analysis on a per-word basis. ● Remove stopwords - Stopwords such as “a”, “an”, “the”, etc. which have very little semantic meaning were removed from the text which helps in reducing noise and focuses on more important words in the analysis. ● Lemmatization/Stemming - this technique was applied to reduce the words into their base form. Lemmatization maps words to their dictionary form while Stemming reduces words to their truncated form by removing prefixes and suffixes. ● Remove emoji symbols - emojis which were present in the text were removed. Even though emojis can convey sentiments, for the purpose of this study, only the textual element of the review was considered. This helped maintain consistency in the textual analysis.

The review text structures are very complex. The traditional machine learning method produces an above average performance. From the below, it can be seen that the accuracy score of the Logistic Regression model was only 0.625.

### Model 2: Fine-Tuning DistillBert model.

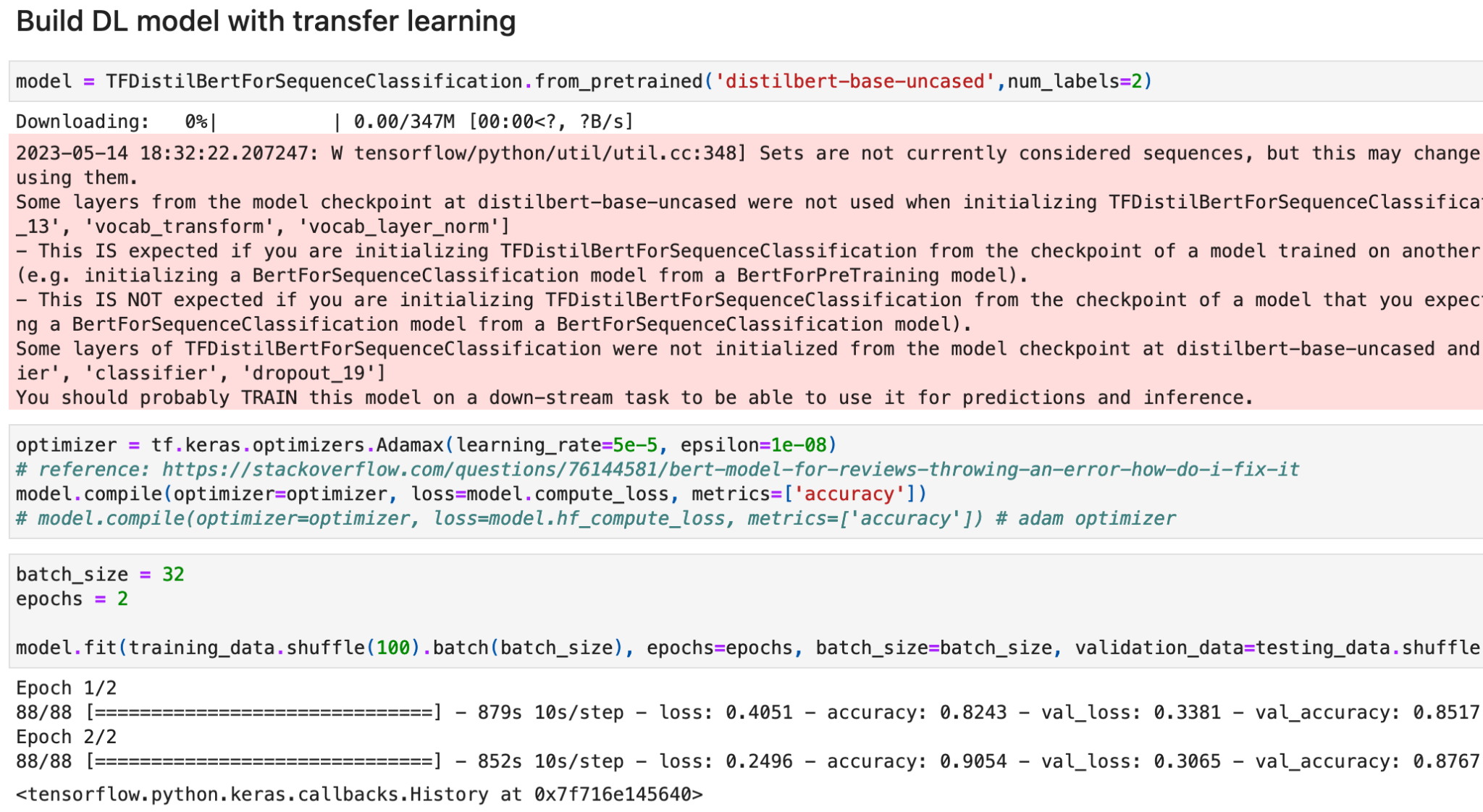
For Model 2, we took full advantage of the pretrained open source transformer model and tensorflow framework. In particular, we used DistillBert for tokenization and for fine-tuning practice to better achieve our own specific task. The tokenization and loading pretrained DistillBert setup are presented below.

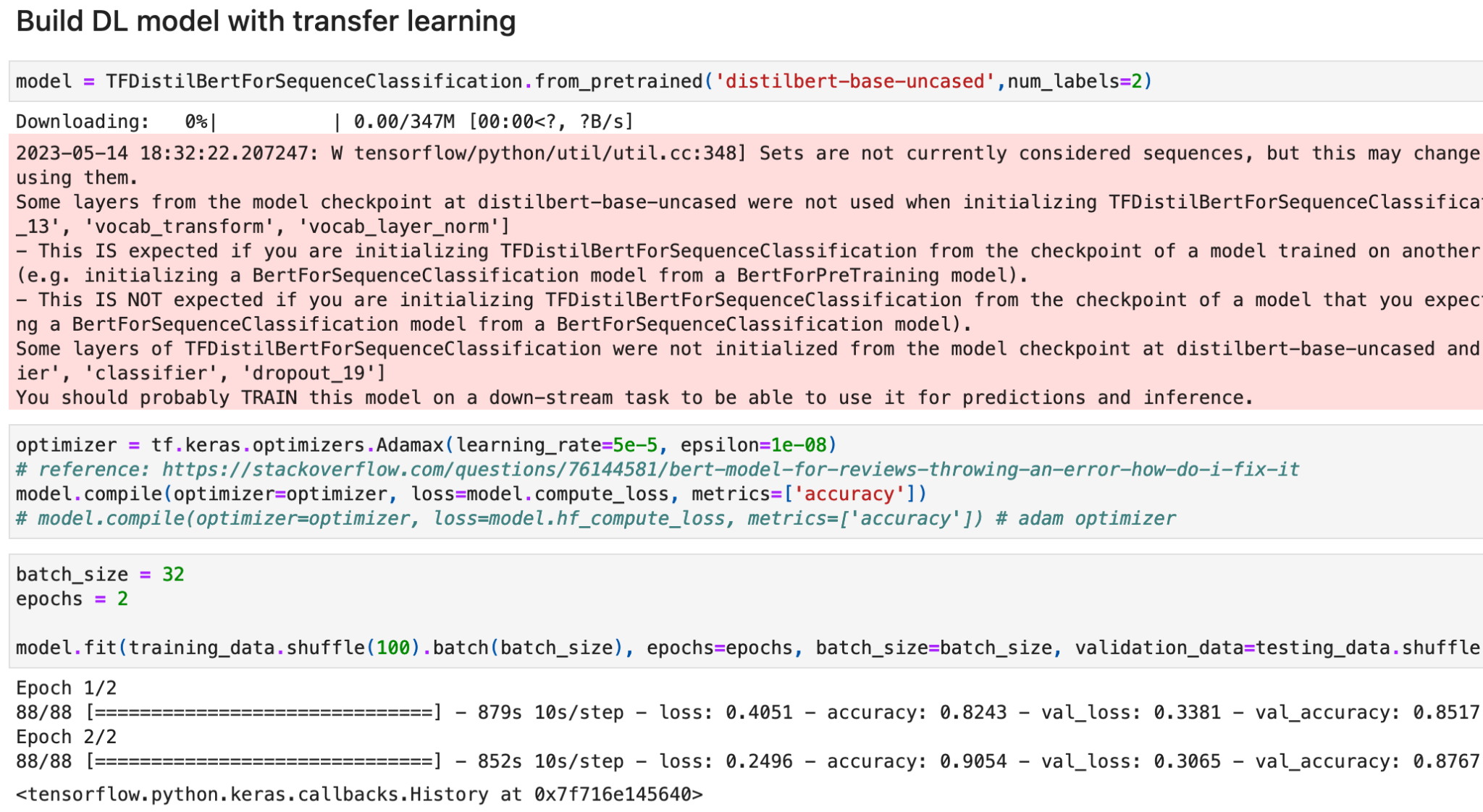
Fine-tuning transformer models raised a few concerns

● Dedicate learning rate on the order of 1E-05

● Fine-tuning process can be computationally expensive.

To this end, we fine-tuned the DistillBert model using two epochs with a learning rate of 5E-05 and Adamax as model optimizer.





The model performed very well with a modest fine-tuning. The main reason is due to the vast source of large corpus used to train transformer models.

Two demonstration tests were presented with our fine-tuned Bert model for illustration purposes on the model performance.



### Integration of GPT-4.

Large Language Models (LLM) are the most trending topic recently. We have explored the surface of leveraging LLM to perform sentiment analysis by providing a crafted prompt text. The prompt we used is as follows:

“‘Perform sentiment analysis on the input delimited by triple quote and return the result in JSON format with two keys. One key is the input review text and another key is classification result, either positive or negative’”

A few trials were conducted. Two test results were recorded using selected sample reviews. The first test showed LLM is on par with the previous generation Transformers model. The second test LLM achieved 100% accuracy. Using LLM in production is still early. More research will be needed.

# Results

In this study, we looked at the Amazon Electronics Review Dataset and performed Sentiment analysis on it using the Logistic Regression Model and DistillBert Model. We also used GPT-4 integration to calculate the sentiment scores of the model.

We then converted the positive label to 1 and negative label to 0 for evaluation. Table 1 below shows the accuracy of each of these models.

Table 1 - Table Showing Models and their Accuracy

| Models | Accuracy Score | Comment |
| --- | --- | --- |
| Logistic Regression Model | 0.62 | Worst Performing Model |
| DistillBert Model | 0.87 | Useable Model |
| GPT-4 Integration | 1.00 | Most accurate model obtained by using a very limited sample size. GPT does show its inconsistency performance |

From the above, it can be observed that the Logistic Regression Model had the lowest accuracy while the GPT-4 had the best accuracy. However, as the GPT-4 integration was applied on a very limited sample size, we have decided to overlook the model. Thus, from this study, the best model is the DistillBert model with an accuracy of 0.87.

# Discussion

One of the main reasons that Logistic Regression performance is shadowed by Distillbert is the way that treatment of negation stopwords during the NLP preprocess stage.

However, they are very important to sentiment analysis. For example, words like ‘not’ and ‘none’ are removed during the remove stopword process. Review texts like “I like this product” and “I don’t like this product” have very different sentiments towards the item that the customer is reviewing. This information is lost. Distillbert models suffered in a similar way when it comes to treating negation stopwords but its performance was compensated by its context preserving capacity.

The above example shows that the DistillBert model has a better contextual understanding of the text and is more effective in capturing the nuanced meaning and semantics of the text which is crucial in sentiment analysis. Similarly, Distillbert is a model which is pretrained on a large corpus of textual data such as Wikipedia or other web document. This pretrained model captures a wide range of language patterns and knowledge making it more suitable for NLP tasks.

Of course, the usage of the appropriate model depends on the task at hand as both Logistic Regression Model and Distillbert Model have their own advantages. Logistic Regression Model can be useful when the problem is relatively simple, and a straightforward relationship between features and sentiment can be expected. LRM also requires less computational resources and training time. Distillbert Model offers more advantages in capturing contextual understanding, leveraging pretrained language representation, and handling long-range dependencies. Thus, the choice between LR and DB model comes down to the requirement of the task, complexity of the problem and the desired balance between performance and interpretability.

# Conclusion & Future work

In conclusion, this study on sentiment analysis of Amazon product reviews using logistic regression, DistillBert, and GPT-4 has provided valuable insights into the effectiveness of different models in capturing sentiment. DistillBert, with its contextual understanding and transfer learning capabilities, outperformed logistic regression in terms of accuracy and capturing nuanced meaning. Additionally, the integration of GPT-4 showcased the potential for advanced language models to enhance sentiment analysis. Further research can explore fine-tuning models, ensemble methods, aspect-based analysis, domain adaptation, and addressing biases to improve sentiment analysis in e-commerce settings.

Moving forward, future research can focus on fine-tuning DistillBert on larger domain-specific datasets to improve performance. Exploring ensemble methods and aspect-based sentiment analysis can enhance accuracy and provide more granular insights. Additionally, domain adaptation techniques, addressing biases, and ensuring fairness are crucial areas to explore for unbiased and inclusive sentiment analysis. By continually refining sentiment analysis techniques, valuable insights from customer feedback can be derived to inform decision-making processes in the dynamic e-commerce landscape.

# References

Aust, F., & Barth, M. (2022). *papaja: Prepare reproducible APA journal articles with R Markdown*. Retrieved from <https://github.com/crsh/papaja>

Barth, M. (2022). *tinylabels: Lightweight variable labels*. Retrieved from <https://cran.r-project.org/package=tinylabels>

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from [https://www.R-project.org/](https://www.r-project.org/)

Analytics Vidhya. (2021, June). Amazon Product Review Sentiment Analysis using BERT. Retrieved from https://www.analyticsvidhya.com/blog/2021/06/amazon-product-review-sentiment-analysis-using-bert/

TensorFlow. (n.d.). Retrieved from https://www.tensorflow.org/

Hugging Face. (n.d.). DistilBERT - Base Uncased. Retrieved from https://huggingface.co/distilbert-base-uncased

Sahoo, S. (2021, May 18). Hugging Face DistilBERT + TensorFlow for Custom Text Classification. Medium. Retrieved from https://medium.com/geekculture/hugging-face-distilbert-tensorflow-for-custom-text-classification-1ad4a49e26a7

Hugging Face. (n.d.). DistilBERT. Retrieved from https://huggingface.co/docs/transformers/model\_doc/distilbert

OpenAI. (n.d.). Retrieved from https://openai.com/

Stack Overflow. (n.d.). Retrieved from <https://stackoverflow.com/>

Oli, B., Zhang, H., & Johar, S.S. Sentiment Analysis on Amazon electronics Dataset. Retrieved from https://www.youtube.com/watch?v=TMhs26prb38